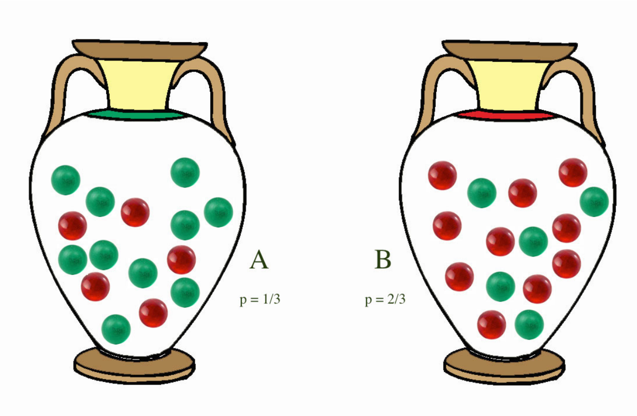
A short introduction to Bayesian statistics, part I

Math 218, Mathematical Statistics
D Joyce, Spring 2016

I'll try to make this introduction to Bayesian statistics clear and short. First we'll look as a specific example, then the general setting, then Bayesian statistics for the Bernoulli process, for the Poisson process, and for normal distributions.

1 A simple example

Suppose we have two identical urns—urn A with 5 red balls and 10 green balls, and urn B with 10 red balls and 5 green balls. We'll select randomly one of the two urns, then sample with replacement that urn to help determine whether we chose A or B.



Before sampling we'll suppose that we have "prior" probabilities of $\frac{1}{2}$, that is, $P(A) = \frac{1}{2}$ and $P(B) = \frac{1}{2}$.

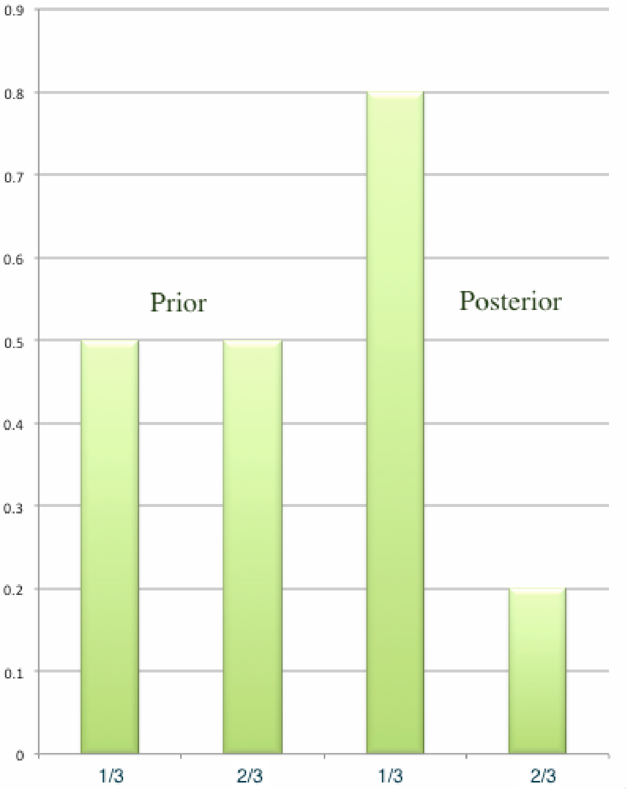
Let's take a sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ of size n, and suppose that k of the n balls we select with replacement are red. We want to use that information to help determine which of the two urns, A or B, we chose. That is, we'll compute $P(A|\mathbf{X})$ and $P(B|\mathbf{X})$. In order to do find those conditional probabilities, we'll use Bayes' formula. We can easily compute the reverse probabilities

$$\begin{aligned} P(\mathbf{X}|A) &= \left(\tfrac{1}{3}\right)^k \left(\tfrac{2}{3}\right)^{n-k} \\ P(\mathbf{X}|B) &= \left(\tfrac{1}{3}\right)^{n-k} \left(\tfrac{2}{3}\right)^k \end{aligned}$$

so by Bayes' formula we derive the posterior probabilities

$$\begin{aligned} P(A|\mathbf{X}) &= \frac{P(\mathbf{X}|A)P(A)}{P(\mathbf{X}|A)P(A) + P(\mathbf{X}|B)P(B)} \\ &= \frac{\left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{n-k} \frac{1}{2}}{\left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{n-k} \frac{1}{2} + \left(\frac{1}{3}\right)^{n-k} \left(\frac{2}{3}\right)^k \frac{1}{2}} \\ &= \frac{2^{n-k}}{2^{n-k} + 2^k} \\ P(B|\mathbf{X}) &= 1 - P(A|\mathbf{X}) \\ &= \frac{2^k}{2^{n-k} + 2^k} \end{aligned}$$

For example, suppose that in $n = 10$ trials we got $k = 4$ red balls. The posterior probabilities would become $P(A|\mathbf{X}) = \frac{4}{5}$ and $P(B|\mathbf{X}) = \frac{1}{5}$.



Before the experiment we chose the two urns each with probability $\frac{1}{2}$, that is, the probability of choosing a red ball was either $p = \frac{1}{3}$ or $p = \frac{2}{3}$ each with probability $\frac{1}{2}$. That's shown in the prior graph on the left. After drawing $n = 10$ balls out of that urn (with replacement) and getting $k = 4$ red balls, we update the probabilities. That's shown in the posterior graph on the right.

How this example generalizes. In the example we had a discrete distribution on p, the probability that we'd chose a red ball. This parameter p could take two values: p could be $\frac{1}{3}$ with probability $\frac{1}{2}$ when we chose urn A, or p could be $\frac{2}{3}$ with probability $\frac{1}{2}$ when we chose urn B. We actually had a prior distribution on the parameter p. After taking into consideration the outcome k of an experiment, we had a different distribution on p. It was a conditional distribution $p|k$.

In general, we won't have only two different values on a parameter, but infinitely many; we'll have a continuous distribution on the parameter instead of a discrete one.

2 The basic principle

The setting for Bayesian statistics is a family of distributions parametrized by one or more parameters along with a prior distribution for those parameters. In the example above we had a Bernoulli process parametrized by one parameter p the probability of success. In the example the prior distribution for p was discrete and had only two values, $\frac{1}{3}$ and $\frac{2}{3}$ each with probability $\frac{1}{2}$.

A sample $\mathbf{X}$ is taken, and a posterior distribution for the parameters is computed.

Let's clarify the situation and introduce terminology and notation in the general case where X is a discrete random variable, and there is only one discrete parameter θ. (In practice, θ is a continuous parameter, but in the example above it was discrete, and for this introduction, let's take θ to be discrete). In statistics, we don't know what the value of θ is; our job is to make inferences about θ. The way to find out about θ is to perform many trials and see what happens, that is, to select a random sample from the distribution, $\mathbf{X} = (X_1, X_2, ..., X_n)$, where each random variable X_i has the given distribution. The actual outcomes that are observed I'll denote $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

Now, different values of θ lead to different probabilities of the outcome $\mathbf{x}$, that is, $P(\mathbf{X}{=}\mathbf{x} \,|\, \theta)$ varies with θ. In the so-called "classical" statistics, this probability is called the *likelihood* of θ given the outcome $\mathbf{x}$, denoted $L(\theta \,|\, \mathbf{x})$. The reason the word likelihood is used is the suggestion that the real value of θ is likely to be one with a higher probability $P(\mathbf{X}{=}\mathbf{x} \,|\, \theta)$. But this likelihood $L(\theta \,|\, \mathbf{x})$ is *not* a probability about θ. (Note that "classical" statistics is much younger than Bayesian statistics and probably should have some other name.)

What Bayesian statistics does is replace this concept of likelihood by a real probability. In order to do that, we'll treat the parameter θ as a random variable rather than an unknown constant. Since it's a random variable, I'll use an uppercase Θ. This random variable Θ itself has a probability mass function, which I'll denote $f_\Theta(\theta) = P(\Theta{=}\theta)$. This f_Θ is called the *prior distribution* on Θ. It's the probability you have *before* considering the information in X, the results of an observation.

The symbol $P(\mathbf{X}{=}\mathbf{x} \,|\, \theta)$ really is a conditional probability now, and it should properly be written $P(\mathbf{X}{=}\mathbf{x} \,|\, \Theta{=}\theta)$, but I'll abbreviate it simply as $P(\mathbf{x} \,|\, \theta)$ and leave out the references to the random variables when the context is clear. Using Bayes' law we can invert this conditional probability. In full, it says

$$P(\Theta{=}\theta \,|\, \mathbf{X}{=}\mathbf{x}) = \frac{P(\mathbf{X}{=}\mathbf{x} \,|\, \Theta{=}\theta) P(\Theta{=}\theta)}{P(\mathbf{X}{=}\mathbf{x})}$$

but we can abbreviate that as

$$P(\theta \,|\, \mathbf{x}) = \frac{P(\mathbf{x} \,|\, \theta) P(\theta)}{P(\mathbf{x})}.$$

This conditional probability $P(\theta \,|\, \mathbf{x})$ is called the *posterior distribution* on Θ. It's the probability you have *after* taking into consideration new information from an observation. Note that the denominator $P(\mathbf{x})$ is a constant, so the last equation says that the posterior distribution $P(\theta \,|\, \mathbf{x})$ is proportional to $P(\mathbf{x} \,|\, \theta)P(\theta)$. I'll write proportions with the traditional symbol $\propto$ so that the last statement can be written as

$$P(\theta \,|\, \mathbf{x}) \propto P(\mathbf{x} \,|\, \theta)P(\theta).$$

Using proportions saves a lot of symbols, and we don't lose any information since the constant of proportionality $P(\mathbf{x})$ is known.

When we discuss the three settings—Bernoulli, Poisson, and normal—the random variable X will be either discrete or continuous, but our parameters will all be continuous, not discrete (unlike the simple example above where our parameter p was discrete and only took the two values $\frac{1}{3}$ and $\frac{2}{3}$). That means we'll be working with probability density functions instead of probability mass functions. In the continuous case there are analogous statements. In particular, analogous to the last statement, we have

$$f(\theta \,|\, \mathbf{x}) \propto f(\mathbf{x} \,|\, \theta)f(\theta)$$

where $f(\theta)$ is the prior density function on the parameter Θ, $f(\theta \,|\, \mathbf{x})$ is the posterior density function on Θ, and $f(\mathbf{x} \,|\, \theta)$ is a conditional probability or a conditional density depending on whether X is a continuous or discrete random variable.

3 The Bernoulli process.

A single trial X for a Bernoulli process, called a Bernoulli trial, ends with one of two outcomes—success where $X = 1$ and failure where $X = 0$. Success occurs with probability p while failure occurs with probability $q = 1 - p$.

The term Bernoulli process is just another name for a random sample from a Bernoulli population. Thus, it consists of repeated independent Bernoulli trials $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ with the same parameter p.

The problem for statistics is determining the value of this parameter p. All we know is that it lies between 0 and 1. We also expect the ratio k/n of the number of successes k to the number trials n to approach p as n approaches ∞, but that's a theoretical result that doesn't say much about what p is when n is small.

Let's see what the Bayesian approach says here. We start with a prior density function $f(p)$ on p, and take a random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Then the posterior density function is proportional to a conditional probability times the prior density function

$$f(p \,|\, \mathbf{x}) \propto P(\mathbf{X}{=}\mathbf{x} \,|\, p)\, f(p).$$

Suppose, now, that there are k successes occur among the n trials $\mathbf{x}$. With our convention that $X_i = 1$ means the trial X_i ended in success, that means that $k = x_1 + x_2 + \cdots + x_n$. Then

$$P(\mathbf{X}{=}\mathbf{x} \,|\, p) = p^k(1-p)^{n-k}.$$

Therefore,

$$f(p \,|\, \mathbf{x}) \propto p^k(1-p)^{n-k}\, f(p).$$

Thus, we have a formula for determining the posterior density function $f(p \,|\, \mathbf{x})$ from the prior density function $f(p)$. (In order to know a density function, it's enough to know what it's proportional to, because we also know the integral of a density function is 1.)

But what should the prior distribution be? That depends on your state of knowledge. You may already have some knowledge about what p might be. But if you don't, maybe the best thing to do is assume that all values of p are equally probable. Let's do that and see what happens.

So, assume now that the prior density function $f(p)$ is uniform on the interval $[0, 1]$. So $f(p) = 1$ on the interval, 0 off it. Then we can determine the posterior density function. On the interval $[0, 1]$,

$$\begin{array}{rcl} f(p \,|\, \mathbf{x}) & \propto & p^k(1-p)^{n-k}\, f(p) \\ & = & p^k(1-p)^{n-k} \end{array}$$

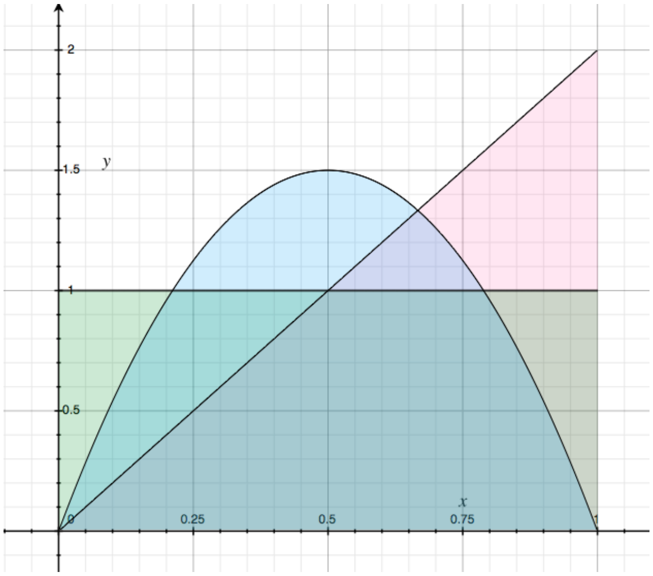
That's enough to tell us this is the beta distribution $\textsc{Beta}(k+1, n+1-k)$ because the probability density function for a beta distribution $\textsc{Beta}(\alpha, \beta)$ is

$$f(x) = \frac{1}{\mathrm{B}(\alpha, \beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$$

for $0 \le x \le 1$, where $\mathrm{B}(\alpha, \beta)$ is a constant, namely, the beta function B evaluated at the arguments α and β.

Note that the prior distribution $f(p)$ we chose was uniform on $[0, 1]$, and that's actually the beta distribution $\textsc{Beta}(1, 1)$.

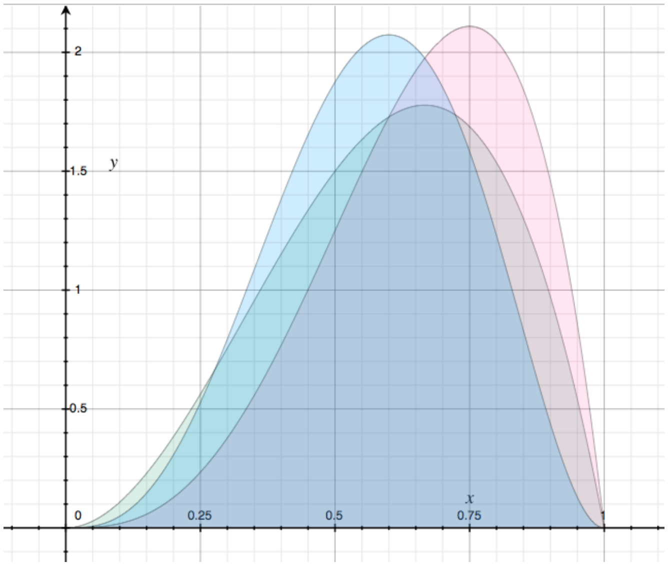
Let's suppose you have a large number of balls in an urn, every one of which is either red or green, but you have no idea how many there are or what the fraction p of red balls there are. They could even be all red or all green. You decide to make your prior distribution on p uniform, that is $\textsc{Beta}(1, 1)$. This uniform prior density is shaded green in the first figure.



Now you choose one ball at random and put it back. If it was red, your new distribution on p is $\textsc{Beta}(2, 1)$. The density function of this distribution is $f_P(p) = 2p$. It's shaded pink in the figure. The probability is now more dense near 1 and less near 0.

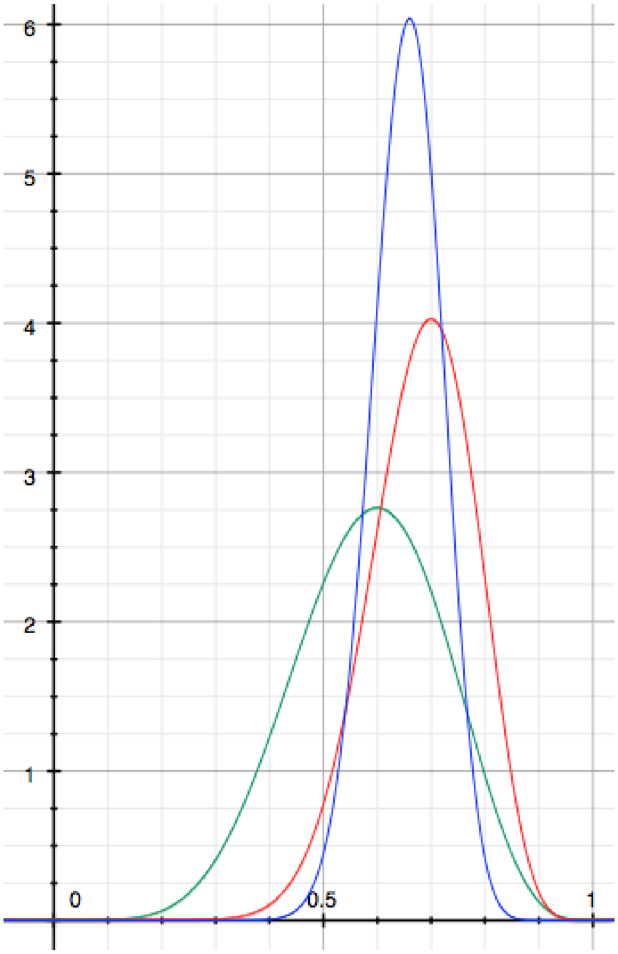
Let's suppose we do it again and get a green ball. Now we've got $\textsc{Beta}(2, 2)$. So far, one red and one green, and the probability is shifted back towards the center. Now $f_P(p) = 6p(1-p)$. It's shaded blue in the figure.

Suppose the next three are red, red, and green, in that order. After each one, we can update the distribution. Next will be $\textsc{Beta}(3, 2)$, then $\textsc{Beta}(4, 2)$, and then $\textsc{Beta}(4, 3)$ They appear in the next figure. The first green, second pink, and third blue.



With each new piece of information the slowly narrows. We can't say much yet with only 5 drawings, 3 reds and 2 greens. A sample of size 5 doesn't say much. Even with so little information, we can still pretty much rule out p being less than 0.05 or greater than 0.99.

Let's see what the distribution would look like with more data. Take three more cases. First, when $n = 10$ and we've drawn red balls 6 times. Then when $n = 20$ and we've gotten 14 red balls. And finally when $n = 50$ and we got 33 reds. Those have the three distributions $\textsc{Beta}(7, 5)$ graphed in green, $\textsc{Beta}(15, 7)$ graphed in red, and $\textsc{Beta}(34, 18)$ graphed in blue. These are much skinnier distributions, so we'll squeeze the vertical scale.



Even after 50 trials, about all we can say is that p is with high probability between 0.4 and 0.85. We can actually compute that high probability as well, since we have a distribution on p.

Math 218 Home Page at
`http://math.clarku.edu/~djoyce/ma218/`